ChinaXiv合作期刊

科技情报分析中 LDA 主题模型最优主题数确定方法研究*

关 鹏[1,2] 王曰芬[1]
[1](南京理工大学经济管理学院 南京 210094)
[2](巢湖学院应用数学学院 合肥 238000)

摘要:【目的】有效确定科技情报分析中 LDA 主题模型的最优主题数目。【方法】利用主题相似度度量潜在主题之间的差异,同时结合困惑度提出一种确定 LDA 最优主题数目的方法,该方法既考虑主题抽取效果同时也考虑模型对新文档的泛化能力。【结果】获取国内新能源领域的科技文献作为数据集,实证结果表明本文提出的最优 LDA 主题数确定方法与单纯使用困惑度相比,具有更高的主题抽取查准率(91.67%)、F 值(86.27%)及科技文献推荐精度(71.25%)。【局限】未针对其他类型的数据集进行新方法的验证,如微博短文本、XML 文档等。【结论】本文方法能够有效地从科技文献数据集中抽取辨识度较高的主题,并能够提高科技文献推荐效果。

关键词: LDA 主题模型 相似度 困惑度 科技情报分析
分类号: G202

chinaXiv:201711.02043v1

1 引 言

LDA(Latent Dirichlet Allocation)[1]主题模型是统计语言模型中的典型代表,近几年在情报分析、知识服务、知识发现等领域得到了广泛的应用,主要集中在科学文献知识挖掘[2-4]、科学研究热点发现与新兴主题探测[5-7]、科学研究主题演化[8-10]、学术评价[11]等研究方向。LDA 之所以在情报学领域获得了广泛的应用,主要原因在于LDA适合海量异构文本数据的建模,其优势是可以将文本表示的维度大大降低,从而避免维数灾难[12]。科技情报分析中大量实证研究证明了 LDA 的可靠性和有效性,但仍存在一些问题没有解决。与一般的文本挖掘任务相比,科技情报分析对LDA提出了更高的要求,主要表现在以下两点:

(1) 在一般的文本挖掘任务中(如文本聚类、文本分类、文本自动摘要[13-16]等), LDA 往往在中间的降维环节发挥重要作用,不需要展示主题的具体形式,只需要实现文本降维即可。但在科技情报分析任务中(如科学研究主题发现与主题演化), LDA 必须将主题抽取的结果展示并分析,主题抽取的质量直接影响主题抽取和主题演化的效果。

(2) LDA 在情报分析中的应用更注重主题数目的确定。目前普遍认为应用 LDA 的最大问题是无法确定最优主题数目[17]。而主题数目的确定对于科技文献主题抽取至关重要。从目前国内外情报学领域应用 LDA 进行科技情报分析的情况看,以上的两个问题还没有引起足够的重视。

2 相关工作

大量实证研究证实 LDA 主题抽取效果与潜在主题数目 K 值有直接关系,主题抽取的结果对 K 值非常敏感。基于此,国内外不少学者展开了相关研究,通过各种

通讯作者: 王曰芬, ORCID: 0000-0002-7143-7766, E-mail: yuefen163@163.com。

*本文系国家自然科学基金研究项目“新研究领域科学文献传播网络生长及对传播效果影响研究”(项目编号: 71373124)、国家社会科学基金重点项目“大数据环境下社会舆情与决策支持方法体系研究”(项目编号: 14AZD084)和江苏高校哲学社会科学重点研究基地(培育点)“社会计算与舆情分析”的研究成果之一。

ChinaXiv合作期刊

chinaXiv:201711.02043v1

方法确定最优主题数目，比较常用方法有以下三种:

(1) Blei 等采用困惑度(Perplexity)作为评价模型好坏的标准，通过选取困惑度最小的模型确定主题的最优数目[1]。困惑度指标可以确定最优的模型预测能力，但是根据困惑度选取的主题数目往往偏大，从而导致抽取的主题之间相似度较大，主题辨识度不高的问题，影响科技情报分析工作的效率。

(2) 将主题数目进行非参数化处理，典型代表是层次狄利克雷过程(Hierarchical Dirichlet Processes, HDP)[18]。HDP 与 LDA 主题模型不同的是: HDP 是一种非参数贝叶斯模型，能够从文档集中自动训练最合适的主题数目 K。HDP 通过狄利克雷过程的非参数特性解决了 LDA 中主题数目选择问题，实验证实 HDP 所选的最优主题数目与基于困惑度选取的最优主题数目一致。但这种方法需要为同一个集合分别建立一个 HDP 模型和一个 LDA 模型，且算法时间复杂性较高，应用在科技情报分析中存在效率不高的问题。

(3) Griffiths 等提出应用贝叶斯模型确定最优主题数目的方法[19]。该方法依赖于 Gibbs 抽样的过程，计算复杂度较高，且只能用来确定主题数目，无法刻画模型的泛化能力。

另外，一些学者探讨了主题相似度和最优主题数目之间的关联。Arun 等将 LDA 看作矩阵分解过程，主题抽取的效果取决于 K 值的选取，并通过实验发现利用 KL 散度度量主题之间的相似度，当主题数接近最优值时，KL 散度较小，而主题数远离最优值时，KL 散度较大[20]。曹娟等通过理论证明和实验分析，得到最优主题数与主题相似度之间的关系。以此为约束条件，将最优 K 值选择与 LDA 模型参数估计统一在一个框架里，通过实验证明最优 K 值不仅与文档集中文本的数量有关，也与文本之间的相关程度有关[21]。综合分析发现，以上确定LDA最优主题数的方法，主要存在模型复杂度较高或者分析所得主题的辨识度不高等问题，基于此，本文从主题相似度入手构建新的确定 LDA 主题数的方法。

3 基于主题相似度和困惑度的最优 LDA 主题数确定方法

如前所述，当使用LDA对科技文献集进行主题抽取时，困惑度选取的主题数目往往较大、从而导致抽取的主题之间相似度较大、主题辨识度不高的问题。而主题辨识度与主题之间的相似度密切关联，当主题相似度越小时，主题之间的辨识度越大。基于此，本文权衡模型的泛化能力以及主题抽取的效果，提出基于困惑度和主题相似度相结合的指标Perplexity-Var来确定主题的最优数目。

3.1 困惑度

在概率语言模型中，困惑度是用来评估语言模型优劣的指标，其基本思想是给测试集赋予较高概率值的语言模型较好[22]，且较小的困惑度意味着模型对新文本有较好的预测作用，所以困惑度一般随着潜在主题数量的增加呈现递减的规律。

在 LDA 主题模型中，困惑度计算公式[1]如下:

$$\text{Perplexity}(D)=\exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \tag{1}$$

其中, D 表示语料库中的测试集，共 M 篇文档, N_d 表示每篇文档 d 中的单词数, w_d 表示文档 d 中的词, $p(w_d)$即文档中词 w_d 产生的概率。

3.2 Perplexity-Var

计算主题相似度常用的方法是 Kullback-Leibler 散度(KL 散度)[23]或 Jensen-Shannon 散度(JS 散度)[24]，由于 KL 散度不满足对称性和三角不等式，所以本文选取 JS 散度作为度量主题之间相似度的计算方法。

在 JS 散度的基础上，将随机变量方差的概念引入到潜在主题空间中，即可衡量主题空间的整体差异性。主题方差是各个主题分别与其均值之间的距离平方和的平均数，用 Var(T)表示。主题方差用来度量主题和其均值之间的偏离程度，可以衡量潜在主题空间的整体差异性和稳定性。主题方差的计算方法如下:

①计算主题–词概率分布 ϕ 均值 $\bar{\phi}$;

②利用 JS 散度计算主题方差，公式如下:

$$\text{Var}(T)=\frac{\sum_{i=1}^{K}[D_{JS}(T_i,\bar{\phi})]^2}{K} \tag{2}$$

其中, T 表示 LDA 抽取的主题, K 表示主题数目, D_{JS} 表示 JS 散度。Var(T)衡量了主题之间的稳定性和差异性，当 Var(T)越大时，主题之间的差异性越大，主题之间的区分性就越好，这样的主题结构就越稳定。困惑度反映了模型的预测能力，但一味追求模型的预测能力则必然导致抽取的主题数过大的问题，所以二者相结合可以有效解决主题辨识度不高的问题。

ChinaXiv合作期刊

chinaXiv:201711.02043v1

Perplexity-Var 指标计算公式如下:

$$\text{Perplexity-Var}(D_{test}) = \frac{\text{Perplexity}(D_{test})}{\text{Var}(T_{test})} \quad (3)$$

其中, D_{test} 为实验文本集中的测试数据集, $\text{Perplexity}(D_{test})$ 为测试数据集的困惑度, $\text{Var}(T_{test})$ 是测试数据集的主题方差。

Perplexity-Var 指标含义: 首先, 考虑到模型的泛化能力, 当 Perplexity 越小时, LDA 的泛化能力越好。其次, 考虑到 LDA 的主题抽取效果, 当主题结构的平均相似度最小时, 对应的 LDA 主题模型最优[21], 而主题结构的平均相似度越小, 则主题之间的差异就越大, 此时主题结构的方差越大。所以当主题方差越大时, LDA 主题抽取的效果越佳, 同时 Perplexity-Var 指标就越小。综合以上分析, 当 Perplexity-Var 指标最小时, 对应的 LDA 主题模型最优。

4 实验过程

4.1 实验数据与数据预处理

(1) 数据检索

实验数据检索自 CNKI, 通过去重、删除不完整数据, 共获得国内新能源领域 1994 年-2000 年 1 018 篇文献, 字段包括标题、作者、机构、摘要和关键词, 不包括全文。将语料库中 10%的文献用作测试集评估模型, 剩下的文献用来训练 LDA 模型。

通过对 1 018 篇科技文献的标题、关键词、摘要等元数据的分析, 笔者统计了文本集的主题及相关统计数据, 经过课题组成员打标签和专家鉴定, 共获得有效主题 27 个, 包含文献 955 篇, 另外还有主题不明确的文献 63 篇, 具体数据如表 1 所示。

表 1 实验文本集主题及文献量

主题	文献量	主题	文献量
太阳能资源	89	风能资源	60
光伏发电	36	风力发电	55
太阳池	11	风力机	48
太阳能空调	10	沼气池	50
太阳灶	18	沼气发酵	30
太阳能电池	15	生物质能	62
太阳能热水器	69	地热资源	63
太阳能集热器	64	地热井、地热田	22
空气取水	8	地热发电	20
氢能	31	热流	14
海洋石油	20	波力发电	12
天然气水合物	62	潮汐能	13
优化设计	15	核能	9
建模、仿真	59	其他	63

(2) 数据预处理

①抽取领域词典、分词

通过 Python 编程获取 1 018 篇原始文献的关键词, 计算词频并获取领域词典。利用 Python 的 jieba[25]分词软件包对原始文献的摘要进行分词, 并将上一步获取的领域词典作为分词组件的用户词典, 以提高分词的效果。

②LDA 主题模型及工具包选择

LDA 主题抽取由基于 Python 语言的机器学习包 gensim[26]实现, Perplexity-Var 指标的计算以及文档相似度的计算也通过 Python 编程实现。

实验环境是一台 Windows 7 旗舰版操作系统、Intel(R) Core(TM) i5-4570 CPU、3.2GHz、4GB 内存的计算机。

4.2 评价指标和实验结果对比分析

确定 LDA 最优主题数目的三种方法中, 基于 HDP 确定 LDA 最优主题数目的方法算法复杂度较高, 而基于 Gibbs 抽样过程中的贝叶斯模型方法无法刻画模型的新文档预测能力。所以, 本文选取最流行的基于困惑度计算的方法作为本文方法的比较对象。实验设计从科技文献主题抽取效果和科技文献相似度推荐效果两个评价指标进行模型评价。

(1) 科技文献主题抽取效果

采用查准率 P (Precision)、查全率 R (Recall)和 F 值(F-Score)进行定量评价。查准率用以评估 LDA 主题抽取的有效主题中正确主题所占的比例, 查全率用以评估 LDA 抽取的正确主题占专家评判的领域研究主题的比例, 而 F 值为二者的调和平均值, 公式如下:

$$P = \frac{T_{correct}}{T_{extract}};\ R = \frac{T_{correct}}{T_{standard}};\ F = \frac{2PR}{(P+R)} \quad (4)$$

其中, $T_{extract}$ 为 LDA 抽取的有效主题的数目; $T_{correct}$ 为有效主题中正确抽取的主题数目, 所谓正确抽取的主题指 LDA 所抽取的主题包含在专家评判的领域研究主题之中; $T_{standard}$ 为通过文献调研和专家评判的领域主题数目。

(2) 科技文献相似度推荐

高质量的科技情报服务应立足于用户需求, 当用

ChinaXiv合作期刊

chinaXiv:201711.02043v1

户在海量科技文献中寻找与自己阅读文献相似度较高的文献时，科技文献相似度推荐就显得尤为迫切，而文献推荐的质量与所抽取主题的质量是直接相关联的。所以，特别选取科技文献相似度推荐效果作为评价最优主题数目选择方法的依据之一。

对训练集语料库实行LDA主题抽取之后，文档可以表示为主题向量空间，其维度比词向量空间的维度小很多。对于测试集的新文档，可以使用训练好的LDA 模型进行主题抽取，并将文档映射到主题空间，在此基础上使用 JS 散度度量新文档与训练集中文档的相似度，完成新文档的相似度推荐工作。

基于文档相似度的文档推荐方法如下:

①在主题数目为 K 时用训练语料库对 LDA 模型进行参数学习;

②对测试集中的文档用训练好的 LDA 进行主题抽取;

③对测试集中的文档根据 JS 散度与训练集中的所有文档进行相似度计算，JS 散度越小则文档越相似，对所有文档进行相似度排名，排名靠前的文档为相似度高的文档。

实验通过打标签的形式，对测试集中的 102 篇文献进行人工标注，标注出训练数据集中与之最相关的前 10 个文献。对每篇测试集文献取其相似度推荐结果中的前 10 篇文献，通过推荐准确率(Recommend Precision)对相似度推荐效果进行对比分析。

假设对于 M 篇测试集中的文献 d_i，在训练数据集中，人工标注的最相关的前 10 篇文献集为 T_i，通过相似度推荐算法得到的推荐结果前 10 篇文献集为 R_i。则该测试集的推荐精度如公式(5)所示，其中#(T_i)表示文献集 T_i 所含文献数量。

$$RP=\frac{\sum_{i=1}^{M}\frac{\#(T_i\cap R_i)}{\#(T_i)}}{M} \tag{5}$$

4.3　实验结果及对比分析

(1) 最优主题数目的确定

实验设定主题数目 K 的取值范围为[10, 200]，取步长为 10 进行 LDA 主题抽取，分别在测试集上计算 Perplexity指标和Perplexity-Var指标，从而确定最优主题数目。

①Perplexity 指标的计算

从图 1 中困惑度的取值来看，当主题数目 K=70 时，LDA 的困惑度指标达到最小，此时最优主题数目为 70。

②Perplexity-Var 指标的计算

利用 JS 散度在测试集中计算不同主题数目 K 情况下，LDA 抽取的潜在主题的方差如图 2 所示。

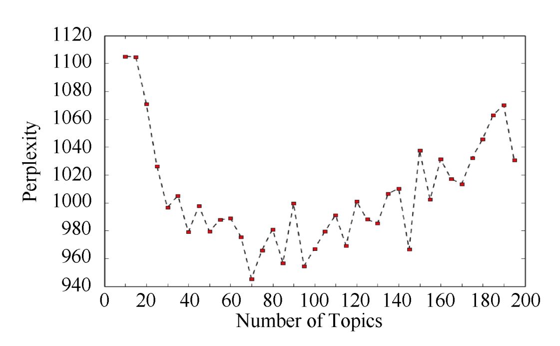


图 1　不同 K 值下 LDA 模型的困惑度

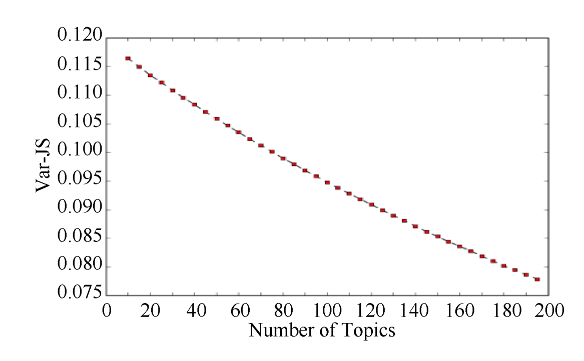


图 2　不同 K 值下 LDA 主题方差值

图 2 中显示方差随着主题数目的增加而减小，即当主题数量越多时，主题之间的方差越小。这是因为当抽取的主题越多时，出现了一些干扰主题和语义重复的主题，导致主题之间的相似度增大，主题结构的方差变小，造成主题结构不稳定。

使用 Perplexity-Var 指标计算最优主题数目，如图 3 所示。可以得出当主题数目选择为 30 个时，Perplexity-Var 指标达到最小值，此时选择的 LDA 最优主题数目为 30。

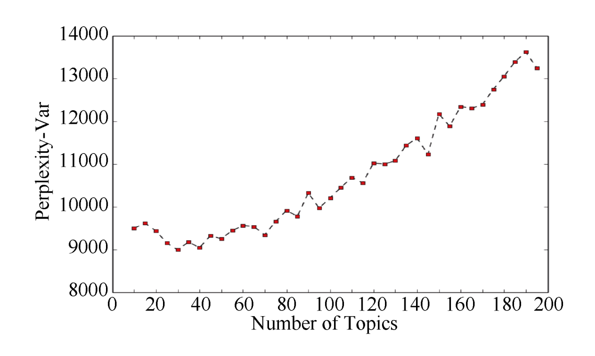


图 3　不同 K 值下 Perplexity-Var 指标值

综上，从两种指标所确定的 LDA 主题数目来看，单纯使用困惑度确定的主题数目 70 与人工判定的主题数目 27 相差太大，而本文所提出的 Perplexity-Var 指标得到的最优主题数目 30 与人工判定的结果比较

吻合。

(2) 实验结果对比分析

①科技文献主题抽取效果评价

根据实验结果可知，通过 Perplexity 指标计算的最优主题数目 70，通过 Perplexity-Var 指标计算的最优主题数目为 30，利用 LDA 对新能源领域科技文献数据集进行主题抽取，并分析结果，部分主题抽取结果如表 2 和表 3 所示(只展示了前 10 个主题并省略了主题词的概率值)：

表 2　K=30 时 LDA 主题抽取结果(部分结果，阴影为干扰主题)

主题	主题词				
Topic1	太阳能热水器	太阳能发电	农村能源	燃气热水器	蓄热
Topic2	太阳能	太阳能集热器	管簇结构腔体式吸收器	集热效率	仿真
Topic3	太阳能集热器	太阳能热水器	保温材料	循环管	聚苯乙烯泡沫板
Topic4	太阳能	设计	发展	热水器	海洋能
Topic5	沼气池	产气量	发酵液	农村	活动盖
Topic6	地温梯度	地热资源	温度	热流	地热场
Topic7	潮汐发电	风力发电机组	间断性发电	温泉水	风机
Topic8	天然气水合物	温室气体	气体水合物	海洋	甲烷
Topic9	太阳灶	反射率	太阳房	太阳能利用	太阳能资源
Topic10	太阳能利用	集热器	太阳能热水器	真空管太阳能热水器	新能源

表 3　K=70 时 LDA 主题抽取结果(部分结果，阴影为干扰主题)

主题	主题词				
Topic1	潮汐电站	潮汐能资源	潮汐能源	灯泡贯流式机组	开发前景
Topic2	太阳能热水器	集热器	热效率	太阳热水器	太阳能干燥器
Topic3	太阳能	热效率	供热与制冷	热损	管簇结构腔体式吸收器
Topic4	地热	厌氧发酵	地热热泵	供暖	太阳能集热器
Topic5	真空集热管	太阳集热器	全玻璃	选择性吸收涂层	真空太阳集热管
Topic6	沼气	综合利用	太阳能资源	天然气	自动绘图
Topic7	地温梯度	大地热流	使用方法	地热电站	瞬时效率
Topic8	风电场	风能	风能资源	风力发电机组	风力机
Topic9	地温场	金属陶瓷	地热	共溅射	太阳能制氢
Topic10	风力机	风力发电机	控制系统	模型	风轮

主题的含义是通过其主题词项的综合语义反映出来的，通过与人工判定的主题进行比较(见表 1)，得出 Perplexity-Var 指标确定的 LDA 主题模型可以准确抽取 22 个主题，所抽取的 30 个主题中含有 6 个干扰主题；Perplexity 指标确定的 LDA 主题模型可以准确抽取 23 个主题，所抽取的 70 个主题中含有 29 个干扰主题。两种指标下的主题抽取效果对比如表 4 所示：

表 4　不同最优主题数选择方法下 LDA 主题抽取效果对比

最优主题数选择方法	$T_{extract}$	$T_{correct}$	$T_{standard}$	查准率(P)	查全率(R)	F 值
Perplexity	41	23	27	56.10%	85.19%	67.65%
Perplexity-Var	24	22	27	91.67%	81.48%	86.27%

表 4 展示了两种最优主题数选择方法下，LDA 主题抽取的查准率、查全率和 F 值。可以看出，基于困惑度(Perplexity)的方法，抽取的有效主题数较多，但是这些主题大多是重复的且干扰主题也很多，所以查准率和 F 值较低。而基于主题相似度和困惑度(Perplexity-Var)的选择方法，抽取的主题中干扰主题较少，各项指标较高，效果较好。科技文献主题挖掘的目标，既要保证主题抽取的准确性也要保证主题抽取有较高的效率。否则，抽取的干扰主题过多，会严重影响主题挖掘效率。

②科技文献相似度推荐

先将训练文本集通过 LDA 进行主题抽取，获取主题空间。然后将测试文本集中的每篇文献表示为主题空间中的向量，利用本文提出的相似度推荐方法推荐相似文献，并取前 10 篇推荐文献。表 5 展示了两种指标下，测试文本集的相似度推荐精度。

chinaXiv:201711.02043v1

ChinaXiv合作期刊

表 5　两种指标下文献相似度推荐精度一览表

最优主题数选择方法	最优主题数目	相似度推荐精度
Perplexity	70	64.76%
Perplexity-Var	30	71.25%

从表 5 看出, Perplexity-Var 指标确定的 LDA 主题模型其文献相似度推荐精度比单纯使用困惑度指标要高，主要原因是 Perplexity-Var 指标不仅依赖于模型的预测能力，还兼顾了主题之间的相似度，使主题之间的差异性更加明显，增加了主题的辨识度。当文档映射到主题空间上时，主题可以很好地表达文档的语义信息。为了更加清晰地展示文献相似度推荐效果，笔者从测试集中随机选取了两篇测试文档进行相似度推荐结果的展示，分别属于潮汐发电主题和风力发电主题，如表 6 和表 7 所示:

表 6　文档相似度推荐结果对比 1

推荐文档(测试集)	K=30 时的推荐结果排序(取前 5)			K=70 时的推荐结果排序(取前 5)		
文档关键词	文档编号(训练集)	JS 散度	文档关键词	文档编号(训练集)	JS 散度	文档关键词
潮汐电站；潮汐能源；潮汐能资源；利用问题；经济效益；电站建设；灯泡贯流式机组；离退休科技工作者；发展前景；开发前景	215	0.00436	潮汐电站；规划设计；浙江省；潮汐能资源；灯泡贯流式机组；潮汐发电站；年发电量；电力负荷；潮汐资源；开发利用	346	0.01760	海洋能资源；开发前景；资源开发利用；波浪能；盐差能；海洋热能；潮汐能资源；潮汐发电站；年发电量；琼州海峡
	299	0.00436	海洋能；可再生能源；潮汐电站；波浪能；开发利用；波浪发电；波力电站；发电装置；装机容量；化石燃料	671	0.01760	地热井；贴砾管；钻机提升系统；过滤器；牙轮钻头；钻井参数；成井工艺；存在问题；测井资料；石油钻井
	346	0.00436	海洋能资源；开发前景；资源开发利用；波浪能；盐差能；海洋热能；潮汐能资源；潮汐发电站；年发电量；琼州海峡	311	0.01760	潮汐能；潮汐电站；综合开发
	444	0.00853	海洋波浪能；波浪能发电站；装机容量；理论蕴藏量；波浪发电；开发利用；年发电量；波能发电站；振荡水柱式；波力电站	406	0.01760	对数正态模型；参数估算方法；拟合误差
	576	0.00853	发电设备；开发利用；波浪能量；发电机；波浪发电；水下波；缩小比；蘑菇形；浮体；样机	459	0.03164	潮汐电站；运行方式；分析

表 7　文档相似度推荐结果对比 2

推荐文档(测试集)	K=30 时的推荐结果排序(取前 5)			K=70 时的推荐结果排序(取前 5)		
文档关键词	文档编号(训练集)	JS 散度	文档关键词	文档编号(训练集)	JS 散度	文档关键词
风力发电场；风力发电机组；风力机；年发电量；雷州半岛；有效风速；总装机容量；风电场；发电装机容量；常规火电	142	0.00073	内蒙古草原；内蒙古锡林浩特；风力发电机；牧民；财政补贴；分离牛奶；小型风机；风能开发；粉碎饲料；风能资源	693	0.10211	风能；风力发电；风电场；现状前景
	196	0.00073	风能资源；有效风能；开发利用前景；风能密度；嵊泗县；有效风速；风力发电；设计风速；相对变率；年平均风速	381	0.10286	风能；风力机；风能利用；风能研究
	214	0.00073	浙江省海岛；有效风速；有效风能密度；风力资源；年平均风速；风能资源；日变化；计算公式；电力紧缺；风资料	607	0.10286	风资源评价；风电场；年平均风速；风能功率密度
	267	0.00191	风力机组；内蒙古锡林浩特；安家落户；风力发电机；锡盟；内蒙古锡林郭勒盟；拖拉机制造厂；西德；电建二公司；年平均风速	752	0.10286	风能；风力发电；装机容量；风电场
	269	0.00209	风电机；内蒙古锡林浩；特大型风力发电机组；风电场；锡林浩特市；风能功率密度；商业化运营；计算机控制；风电机组；拖拉机制造	859	0.10531	可持续发展；风能；风电场；租赁

(注：表 6 和表 7 中 JS 散度表示文档之间的距离，当两篇文档其 JS 散度越小时，二者之间的相似度就越大。)

chinaXiv:201711.02043v1

ChinaXiv合作期刊

从表 6 可知，第一篇属于潮汐发电主题的被推荐文档在主题数目 K=30 时与训练集中文档 215 之间具有最小的 JS 散度，因而最相似；而当 K=70 时，与文档 346 最相似。从文档的关键词可以看出，文档 215 在关键词上与被推荐文档极为相似，都包含“潮汐电站；潮汐能源；潮汐能资源；灯泡贯流式机组；开发；利用”等词，特别是核心词汇“灯泡贯流式机组”，而文档 346 没有。另外，文档 671 是“地热”主题，与“潮汐发电”无关，可见 K=30 时文档推荐效果要优于 K=70。同样的对比方法，从表 7 中也可以得出类似的结论。可见，基于 Perplexity-Var 指标选择的 LDA 模型，由于保证了所抽取的主题结构的稳定性，当文档表示为主题的混合分布时，能够较准确地刻画文档的语义信息，从而在文档相似度推荐方面有更好的表现。

5 结 语

在大数据背景下，对于智能情报分析需求的日益增强，对于能够处理海量文本数据的智能算法的需求日益增强。本文从 LDA 的特点入手，分析了情报分析与一般的文本挖掘中应用 LDA 的主要区别。提出了在情报分析工作中应用 LDA 必须要重视主题抽取的效果和主题数目这两个问题。结合主题相似度以及困惑度，本文提出确定最优主题数目的方法，实证证实了在科技文献的知识挖掘中，利用此方法可以有效确定主题数目获得较好的主题抽取结果，帮助情报分析工作者从海量科技文献中抽取显著主题，并能够提高基于相似度的科技文献推荐效果。

本文在实证分析时针对科技文献数据进行了方法有效性验证，没有针对其他类型的数据集进行方法的验证，如微博短文本、XML 文档等。另外，只针对科技情报分析任务，从主题抽取效果和科技文献相似度推荐效果这两个方面进行新方法的验证，其他方面的验证还需要进一步的拓展，以证明方法的有效性。所以，扩展验证范围和评价指标是下一步的工作重点。

参考文献：

[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] 王萍. 基于概率主题模型的文献知识挖掘[J]. 情报学报, 2011, 30(6): 583-590. (Wang Ping. Literature Knowledge Mining Based on Probabilistic Topic Model [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(6): 583-590.)

[3] Hassan S U, Haddawy P. Analyzing Knowledge Flows of Scientific Literature Through Semantic Links: A Case Study in the Field of Energy [J]. Scientometrics, 2015, 103(1): 33-46.

[4] Liang H, Fang L. Topic Discovery and Trend Analysis in Scientific Literature Based on Topic Model [J]. Journal of Chinese Information Processing, 2012, 26(2): 109-115.

[5] 范云满, 马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究[J]. 情报学报, 2014, 33(7): 698-711. (Fan Yunman, Ma Jianxia. Detection of Emerging Topics Based on LDA and Feature Analysis of Emerging Topics [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(7): 698-711.)

[6] He Q, Chen B, Pei J, et al. Detecting Topic Evolution in Scientific Literature: How Can Citations Help? [C]. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, 2009: 957-966.

[7] AlSumait L, Barbará D, Domeniconi C. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking [C]. In: Proceedings of the 8th IEEE International Conference on Data Mining. 2008.

[8] 刘彤, 杨冠灿, 蒋继娅, 等.基于多重关系的专利网络演化特征与动态分析——以锂离子电池领域为例[J]. 情报学报, 2014, 33(12): 1288-1301. (Liu Tong, Yang Guancan, Jiang Jiya, et al. Research on the Evolution and Dynamic Analysis of Multi-relation Integrated Patent Network: A Case Study on Lithiumion Battery [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(12): 1288-1301.)

[9] 贺亮, 李芳. 科技文献话题演化研究[J]. 现代图书情报技术, 2012(4): 61-67. (He Liang, Li Fang. Topic Evolution in Scientific Literature [J]. New Technology of Library and Information Service, 2012(4): 61-67.)

[10] Wu Q Q, Zhang C D, Hong Q Q, et al. Topic Evolution Based on LDA and HMM and Its Application in Stem Cell Research [J]. Journal of Information Science, 2014, 40(5): 611-620.

[11] Gerrish S, Blei D M. A Language-based Approach to Measuring Scholarly Impact [C]. In: Proceedings of the 27th International Conference on Machine Learning. 2010.

[12] Dhillon I S, Modha D S. Concept Decompositions for Large Sparse Text Data Using Clustering [J]. Machine Learning, 2001, 42(1-2): 143-175.

[13] 王李冬, 魏宝刚, 袁杰. 基于概率主题模型的文档聚类[J]. 电子学报, 2012, 40(11): 2346-2350. (Wang Lidong, Wei

chinaXiv:201711.02043v1

ChinaXiv合作期刊

Baogang, Yuan Jie. Document Clustering Based on Probabilistic Topic Model [J]. Acta Electronica Sinica, 2012, 40(11): 2346-2350.)

[14] Lee H, Kihm J, Choo J, et al. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling [J]. Computer Graphics Forum, 2012, 31(3): 1155-1164.

[15] Kabán A, Girolami M A. A Dynamic Probabilistic Model to Visualise Topic Evolution in Text Streams [J]. Journal of Intelligent Information Systems, 2002, 18(2-3): 107-125.

[16] Chua F C T, Lauw H W, Lim E P. Generative Models for Item Adoptions Using Social Correlation [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(9): 2036-2048.

[17] 张晗，徐硕，乔晓东，等. 融合科技文献内外部特征的主题模型发展综述[J]. 情报学报，2014，33(10)：1108-1120. (Zhang Han, Xu Shuo, Qiao Xiaodong, et al. Review on Topic Models Integrating Intra- and Extra- Features of Scientific and Technical Literature [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(10): 1108-1120.)

[18] Teh Y, Jordan M, Beal M, et al. Hierarchical Dirichlet Processes [J]. Journal of the American Statistical Association, 2007, 101(476): 1566-1581.

[19] Griffiths T L, Steyvers M. Finding Scientific Topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(S1): 5228-5235.

[20] Arun R, Suresh V, Veni Madhavan C E, et al. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations [A]. //Advances in Knowledge Discovery and Data Mining [M]. Springer Berlin Heidelberg, 2010.

[21] 曹娟，张勇东，李锦涛，等.一种基于密度的自适应最优LDA 模型选择方法[J]. 计算机学报，2008，31(10)：1780-1787. (Cao Juan, Zhang Yongdong, Li Jintao, et al. A Method of Adaptively Selecting Bast LDA Model Based on Density [J]. Chinese Journal of Computers, 2008, 31(10): 1780-1787.)

[22] Grossman D A. Information Retrieval: Algorithms and Heuristics [M]. Springer Science & Business Media, 2004.

[23] Duda R O, Hart P E, Stork D G. Pattern Classification [M]. John Wiley & Sons, 2012.

[24] Lin J. Divergence Measures Based on Shannon Entropy [J]. IEEE Transactions on Information Theory, 199l, 37(1): 145-151.

[25] Sun J Y. jieba0.37 [EB/OL]. [2015-10-08]. https://pypi.python.org/pypi/jieba/.

[26] RehurekR. gensim 0.10.2 [EB/OL]. [2014-12-11]. https://pypi.python.org/pypi/gensim.

作者贡献声明:

关鹏：提出研究方案和思路，进行实验，起草并修改论文；

王曰芬：扩展研究思路，审阅论文，提出论文修改建议。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 关鹏. new_energy_corpus (training set).xlsx. 分词、去重、去停用词后的 LDA 训练语料库.

[2] 关鹏. new_energy_corpus (test set).xlsx. 分词、去重、去停用词后的 LDA 测试语料库.

收稿日期: 2016-02-22
收修改稿日期: 2016-03-20

chinaXiv:201711.02043v1

ChinaXiv合作期刊

Identifying Optimal Topic Numbers from Sci-Tech Information with LDA Model

Guan Peng[1, 2] Wang Yuefen[1]
[1](School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094, China)
[2](College of Applied Mathematics, Chaohu University, Hefei 238000, China)

Abstract: [**Objective**] This paper tries to identify the optimal number of topics for the Latent Dirichlet Allocation (LDA) model to analyze scientific and technical information. [**Methods**] First, we used the topic similarity to measure the differences among the latent topics. Second, we proposed a method determining the optimal topic numbers and tried to utilize this model to documents from Chinese literature in the field of new energy. [**Results**] The proposed method achieved higher precision ratio and higher F-score in topic extration, which improved the performance of literature recommendation systems. [**Limitations**] We did not examine the new mothod with other datasets, such as microblog posts and XML documents. [**Conclusions**] The proposed method could identify more recognizable topics and improve the performance of scientific and technical literature recommendation systems.

Keywords: LDA Topic model Similarity Perplexity Analysis of Scientific and Technical Information

chinaXiv:201711.02043v1

开放图书馆基金会成立，旨在促进图书馆开源项目的发展

开放图书馆基金会于近日成立，旨在促进图书馆开源项目的发展，并促进和支持这些开源项目的社区贡献和可持续发展。该基金会为图书馆员、开发人员、设计人员、服务提供商和供应商提供了能够与创新的开源技术进行合作，为图书馆开发转型解决方案的基础架构。

基金会的创建是受到了 FOLIO 项目的启发。FOLIO 项目于 2016 年 6 月启动，到现在，成功创建了一个由图书馆、供应商和软件开发商组成的多元社区，FOLIO 项目的目标是创建一个开源的图书馆服务平台，能够将创新方法运用于现行做法，并鼓励新的和扩展的图书馆服务更全面地支持学术探究和知识生产。该基金会的首届项目包括两个现有的开源社区：开放图书馆环境(Open Library Environment, OLE)和全球开放知识库(Global Open Knowledgebase, GOKb)。OLE 和 GOKb 社区加入开放图书馆基金会因为该基金会专注于图书馆、图书馆社区，以及开放技术和数据。OLE 总经理兼开放图书馆基金 Michael Winkler 指出 OLE 和开放图书馆基金有着共同的目标。“基金会的使命是培育和支持开源项目，这与 OLE 的愿景是一致的。” OLE 正在开发人员、专家和社区基础设施等方面帮助 FOLIO 项目建设 FOLIO 社区。

开放图书馆基金会将确保开源项目所开发的代码的可用性，并且作为这些项目的“避风港”，不受任何贡献者、用户或关联方需求和目标的影响。开放图书馆基金会也将确保代码是遵循 Apache v2 协议免费提供的。欲了解更多信息，请访问 http://www.openlibraryfoundation.org。

(编译自: http://librarytechnology.org/news/pr.pl?id=21867)

(本刊讯)